Toward Interlinking Asian Resources Effectively: Chinese to Korean Frequency-Based Machine Translation System

Eun Ji Kim and Mun Yong Yi(✉)

Department of Knowledge Service Engineering, KAIST, Daejeon, Republic of Korea
{eunjik, munyi}@kaist.ac.kr

Abstract. Interlinking Asian resources on the Web is a significant, but mostly unexplored and undeveloped task. Toward the goal of interlinking Asian online resources effectively, we propose a novel method that links Chinese and Korean resources together on the basis of a new machine translation system, which is built upon a frequency-based model operated through the Google Ngram Viewer. The study results show that Chinese characters can be mapped to corresponding Korean characters with the average accuracy of 73.1 %. This research is differentiated from the extant research by focusing on the Chinese pronunciation system called Pinyin. The proposed approach is directly applicable to voice translation applications as well as textual translations applications.

Keywords: LOD · Asian resources · Machine translation · Google Ngram Viewer · Multilingual resources

1 Introduction

The current Web provides seriously limited support for sharing and interlinking online resources at the data level. Linked Open Data (LOD) is an international endeavor to overcome this limitation of the current Web and create the Web of Data on a global level. The Web of Data is an absolute prerequisite for the realization semantic Web. Since Web 2.0 emerged, a vast amount of data has been released to the LOD cloud in the structured-data format so that computers can understand and interlink them. Many tools and frameworks have been developed to support that transition and successfully deployed in a wide number of areas. However, as the proportion of non-Western data is comparatively small, a couple of projects have been only recently initiated to extend the boundary of LOD over non-Western language resources [1].

Toward the goal of interlinking Asian resources effectively, we pay our first attention to interlinking Chinese and Korean resources and propose a novel method that links Chinese and Korean resources together. Due to the two countries' historical backgrounds, almost all of the Chinese characters can be converted to one or several Korean characters and these Korean characters often reflect the pronunciation of the Chinese characters (e.g., for '夢(dream)' in Chinese, whose pronunciation is 'meng',

W. Kim et al. (Eds.): JIST 2013, LNCS 8388, pp. 69–74, 2014.
DOI: 10.1007/978-3-319-06826-8_6,

it can be uniquely converted to Korean character ‘ 몽’ pronounced as ‘mong’) [2]. According to a study on the frequency in Korean vocabulary use [3], about 70 % of Korean words are common in Chinese and Korean.

In this research, we propose a new method that recognizes the identical or similar nature of the words even though their pronunciations are different, and test its effectiveness using a sample data set consisting of 33,451 words, which is more than enough to cover most of the words commonly used by two countries.

2 Background

Pinyin is an official phonetic system for transcribing the sound of Chinese characters into Latin script in China, Taiwan, and Singapore. It has been used to teach standard Chinese and spell Chinese names in foreign publications and used as an input method to enter Chinese characters into computers [4].

Besides, the usage of Pinyin has come into the spotlight recently in light of the developments in voice recognition technology. It is a good means to express Chinese pronunciation in comparison with Chinese characters, which are not based on phonogram. Furthermore, the total number of Pinyin is about 400 and it is much fewer than the number of Chinese characters.

However, mapping one Pinyin to one Korean pronunciation is not a simple problem. One Chinese character can be read by various Chinese pronunciations and different Chinese characters can have same Pinyin. For example, a Chinese character ‘ 的’ has various Pinyin such as ‘de’ and ‘di’. Also, Chinese characters ‘ 十’ and ‘ 事’ have same Pinyin ‘shi’, but the ‘ 十’ is read as ‘ 십’(sib), and ‘ 事’ is read as ‘ 사’(sa) in Korean.

Therefore, mapping Pinyin to Korean cannot be characterized as 1:1 relationships. As shown in Table 1, we can see that even the same Pinyin can be corresponded to different Korean Characters.

3 Methods

3.1 Resource Used in the Machine Translation System

In this research, we used 3,500 Chinese characters in common use. According to a statistical analysis, 2,400 Chinese characters cover 99 %, 3,800 characters cover 99.9 % of documents written in Chinese [5]. For this reason, we chose 3,500 Chinese characters in common use sets as our standard Chinese characters, covering over 99.7 % (when it was estimated with simple extrapolation).

Table 1. Pinyin to Korean pronunciation relations

Chinese Pronunciation	Korean Pronunciation
Kua	과(gua)
Kuai	괴(gue), 쾌(que), 회(hue)
Kuang	광(gwang), 황(hwang)

3.2 Frequency-Based Table

Utilizing a Pinyin to Korean relation table is unlikely successful because one Pinyin can be mapped to many Korean. Thus, to identify a matching Korean character, we devised a frequency based recommendation system. According to the Chinese Ministry of Education, just 581 Chinese characters can cover more than 80 % of commonly used Chinese language. Exploiting this distribution information, we assigned high points to frequently used Chinese characters so as to properly determine the ranking of corresponding Korean Characters.

For the measurement of frequency in language, we took the advantage of Google Ngram Viewer. Google Ngram Viewer shows statistical data that describe how frequently the queried word was used in books. We queried all 3,500 common Chinese characters and got each character's frequency. A part of the results is shown in the Table 2.

In setting up the frequency measurement, we had to consider when one Chinese character is pronounced in many different ways. In this case, we divided the measured frequency *f* into the number of ways pronounced differently *n* and applied the normalized frequency *f*/*n* to each pronunciation. For example, 的 can be pronounced as 'de' and 'di' in Chinese. Thus, *n* is equal to 2, so the normalized frequency 0.0629/2 was applied to each.

3.3 Pinyin-to-Korean Mapping Table

Through the previous steps, we recorded each Chinese character's frequency. In this section, we explain how to build a Pinyin-to-Korean frequency based translation system. To find their mapping relation, calculating each Korean character's possibility to be mapped to each Pinyin is required. We performed this process using the Eq. (1).

$$Frequency(KP|Pinyin) = \sum_{i=1}^{n} Frequency(kp_i|Pinyin) \quad (1)$$

$Frequency(KP|Pinyin)$ means how frequently a certain Pinyin is chosen by a certain Korean pronunciation *KP* in case of Chinese-Korean common words. Also, kp_i means all of the same Korean pronunciations that came from different Chinese characters. Based on the result of calculating total sum points, Table 3 shows Pinyin to Korean frequency based mapping table.

The way to use this table as a translator is as follows. When the Pinyin input 'a' is given, we can find Korean characters corresponding to 'a' in the table. In this case, both '아' and '가' are candidates. However, the table shows that each candidate's frequency value is different. According to the table, the pronunciation 'a' in Chinese is

Table 2. Frequency of Chinese characters

Chinese Character	Frequency
的	0.0629
是	0.0109
在	0.0106

Table 3. Frequency-based Pinyin to Korean relation

	Pinyin	Mapped Korean Pronunciation			
	a	ㅇ아(ah)	ㄱ가(ga)		
$Frequency(KP\|Pinyin)$		2.06E-5	1.78E-7		
	ai	ㅇ애(eh)	와(wai)		
$Frequency(KP\|Pinyin)$		1.24E-5	3.73E-8		
	an	안(an)	엄(um)	전(jeon)	암(am)
$Frequency(KP\|Pinyin)$		6.33E-5	4.35E-5	9.69E-9	1.78E-8

mapped to Korean ‘ ㅇ아’ with 2.06E-5 frequency while ‘a’ is mapped to Korean ‘ ㄱ가’ with 1.78E-7 frequency. Therefore, this system recommends ‘ ㅇ아’ as the first choice and recommends ‘ ㄱ가’ as the next choice.

4 Experiment

4.1 Experimental Setup

To measure our system’s accuracy, we collected Chinese-Korean common 33,451 words. When Pinyin is given as an input, our Pinyin to Korean translator is executed and it recommends its Korean counterpart. With each recommendation, if it is exactly the same with the original answer we have, it stops to recommend other Korean character and we measure its accuracy according to the number of times it has recommended.

4.2 Performance Evaluation

In this experiment, relying on an existing well known evaluation method was inappropriate because we had to evaluate the accuracy of Korean characters that has not been actively researched and far different from Roman characters. Therefore, we devised our own evaluation method based on a mathematical model.

Our evaluation method measures the recommended Korean’s accuracy compared with the original Korean answer with the score from 0 to 1. In consideration of each word’s length difference, when the length of the word is m, we distributed $1/m$ as a total score to each character. To sum up, each character can take $1/m$ score as its maximum, and the total score for each word cannot exceed 1. In addition, each character’s score is measured by $(1/m) \times (1/j)$ when j means the total number that a new Korean character was recommended. For example, when a firstly recommended character was exactly the same with the compared answer character, j is equal to 1 and the total score for the character becomes $1/m$. However, according to our translation system, if the recommended character is not equal to the answer, it recommends the next ordered character continuously until it can find exactly the same one. For this reason, when the recommended character is low-ranked in its recommendation list, the corresponding character score becomes low. Besides, when there is no recommended character in the recommendation list in the end, the character takes the score of 0.

Table 4. Sample program results for Chinese to Korean machine translation

Chinese	Pinyin	Korean	Recommended Korean	Accuracy
加工	jia gong	가공	가|공|	1
加工工場	jia gong gong chang	가공공장	가|공|공|장|	1
加工順序	jia gong shun xu	가공순서	가|공|순|수{허서}|	0.833333
加工業	jia gong ye	가공업	가|공|야{업}|	0.833333
加工特性	jia gong te xing	가공특성	가|공|특|성|	1

4.3 Evaluation Results

According to the experiment result obtained by using the Chinese-Korean 33,451 common words, a Pinyin input was able to be mapped to a corresponding Korean character with the average accuracy of 73.1 %. Also, the number of words which got 1 in accuracy score was 7,926 and its proportion was about 23.7 %. The machine translation program offered many good translations but still makes errors in recommendation order of priority. For example, the Chinese word ‘ 加工業’(jia gong ye) could not get the score of 1 because the program regarded ‘ 야’(ya) as a corresponding Korean character to ‘ 業’(ye) instead of the correct answer ‘ 업’(up) based on statistical data analysis. The sample results of the translation program are shown in Table 4.

5 Conclusion

In this study, we proposed a new method to translate Chinese to Korean. Different from the extant approaches, we focused on Chinese phonetic system Pinyin. In analyzing 3,500 Chinese characters in common use, we noticed the problem of one to many mapping between the two character systems and consequently realized the necessity of priority in translated results. Thus, we adopted a frequency value which was obtained through Google Ngram Viewer as its ranking criteria. To verify this translator’s performance, we employed about 33,000 words which are commonly used in both Chinese and Korean and evaluated the translated results. Consequently, our Pinyin to Korean frequency based translator showed 73.1 % accuracy on average.

The results of this study have practical implications for linking Asian resources on Web. There are a number of well-developed translator systems for Western languages, whereas translator systems for Asian languages are almost non-existent. However, according to the statistical data of Miniwatts Marketing Group, the percentage of Internet users who are from Asian is bigger than any other continent and its rate is increasing rapidly [6]. Therefore, studies for linking Asian resources will become more important in the near future as there is great necessity for resource sharing among Asians. With this new machine translation system, we can link Chinese resources to Korean resources available in the LOD cloud.

Furthermore, voice translation has become a promising field recently. Google has collected a massive amount of voice data through voice search service. The amount of data Google receives in a day is the same with the amount that a person speaks continuously in two years [7]. In this situation, a demand for research about Pinyin

will be naturally increased. Furthermore, different from most existing machine translation systems, we did not use dictionary definition at all. We devised a rule that exploits the two languages' similarity and applied the rule to automatic translation system. Therefore, the proposed approach requires less time and less memory.

On the other hand, there still remain some improvements that need to be made. Above all, we did not consider the five tones in Chinese language in this research. When we cover five pitches in Chinese, it is expected to have higher accuracy. Besides, the proposed approach can be used for only Chinese-Korean common words. Although a high percentage of Korean words are commonly used in Chinese and Korean, the existence of uncovered words needs to be addressed in the following researches by complementing the proposed approach with other solutions. Notwithstanding these limitations, however, it should be noticed that the proposed approach is a promising first attempt toward interlinking Asian resources effectively.

Acknowledgements. This work was supported by the IT R&D program of MSIP/KEIT. [10044494, WiseKB: Big data based self-evolving knowledge base and reasoning platform]

References

1. Hong, S.G., Jang, S., Chung, Y.H., Yi, M.Y.: Interlinking Korean resources on the web. In: Takeda, H., Qu, Y., Mizoguchi, R., Kitamura, Y. (eds.) Semantic Technology. LNCS, vol. 7774, pp. 382–387. Springer, Heidelberg (2013)
2. Huang, J.-X., Choi, K.-S.: Chinese-Korean word alignment based on linguistic comparison. In: ACL'00 Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pp. 392–399 (2000)
3. Lee, Y.: Research About Korean Chinese Common Words. Samyoung-sa, Seoul (1974)
4. Chen, Z., Lee, K.-F.: A new statistical approach to chinese pinyin input. In: ACL'00 Proceedings of 38th Annual Meeting on Association for Computational Linguistics, pp. 241–247 (2000)
5. Gillam, R.: Unicode Demystified : A Practical Programmer's Guide to the Encoding Standard. Addison-Wesley, Boston (2003)
6. Internet World Stats. http://www.internetworldstats.com/
7. Na, S.H., Jung, H.Y., Yang, S.I., Kim, C.H., Kim, Y.K.: Big data for speech and language processing. Electronics and Telecommunications Trends, pp 52–61 (2013)